Improving Scalability and Usability of Parallel Runtime Environments for High Availability and High Performance Systems

Thara Angskun

Advisor: Dr. Jack Dongarra
Committees: Dr. George Bosilca, Dr. Hairong Qi and Dr. Brad Vander Zanden
Dissertation Statement

- Analyze, understand and improve state of the art mechanisms for managing highly dynamic large-scale applications

- Use new scalable and fault-tolerant topologies combined with rerouting techniques to build parallel runtime environments, which are able to efficiently and reliably deliver sets of information to a large number of processes.
Contributions

- Communication framework (library)
  - MPI implementation independent
  - New logical topologies
    - Graph theoretical (properties) analysis
      - Scalability and Fault-tolerance
    - Fault-tolerant routing protocols
      - Formally specified and verified protocols
    - Performance and reliability analysis
Assumptions

- **External Directory Service**
  - Used in initialization and recovery steps

- **Failures**
  - Assume Fail-stop (rather than Byzantine)

- **Transmission channel**
  - Detect and recover from a transmission error
    - E.g. TCP, Reliable UDP
    - Consequence: never lose a message
Related Works

• Parallel RTE
  • Grid Middleware, SSI, Distributed OS, MPI RTE

• MPI RTE
  • MPD, Open RTE, HARNESS, LAM

• Existing topologies
  • Fully-Connected, Ring, Hypercube, 2D Torus, Hypertree, HiC etc.

• Graph theory
  • (Recursive) Circulant, Knodel, Kautz graphs

• Peer-to-Peer
  • CAN, Chord, SkipNet, Pastry, Tapestry

• Sensor / Ad-hoc
  • Gossiping techniques
Topology Requirements

- Scalable and fault-tolerant capabilities
- No number of node restriction

**Complete K-ary Sibling Tree (CST)**

**Binomial Graph (BMG)**
Complete $K$-ary Sibling Tree (CST)

- $K$ is number of fanout ($k \geq 2$)
Complete $K$-ary Sibling Tree (CST)

- Self-recovery structure
  - Prevent network bisection

- Property analysis
  - Good scalability and fault-tolerance
Complete $K$-ary Sibling Tree (CST)

- Fault-tolerant routing protocol
  - Unicast (1 to 1), multicast (1 to $m$, $m<n$), broadcast (1 to $n$)
- Performance and reliability analysis
  - 50+ experimental evaluations
Complete $K$-ary Sibling Tree (CST)

- Potential Problems

“When I grow up, I want to be a root node.”
Binomial Graph (BMG)

- Undirected graph $G:=(V, E), |V|=n$ (any size)
  - Node $i\in\{0,1,2,\ldots,n-1\}$ has links to a set of nodes $U$
    - $U=\{i\pm1, i\pm2,\ldots, i\pm2^k \mid 2^k \leq n\}$ in a circular space
    - $U=\{(i+1)\mod n, (i+2)\mod n,\ldots, (i+2^k)\mod n \mid 2^k \leq n\}$ and
    - $\{(n+i-1)\mod n, (n+i-2)\mod n,\ldots, (n+i-2^k)\mod n \mid 2^k \leq n\}$
Binomial Graph (BMG)

- Merging all necessary links creates binomial tree from each node in the graph.

Broadcast messages from any node within $\lceil \log_2(n) \rceil$ steps
Theoretical Properties of BMG

- Degree $\delta$ (number of neighbors)

  \[
  \delta = \begin{cases} 
  (2 \times \lfloor \log_2 n \rfloor) - 1 & \text{For } n = 2^k, \text{where } k \in \mathbb{N} \\
  (2 \times \lfloor \log_2 n \rfloor) - 2 & \text{For } n = 2^k + 2^j, \text{where } k, j \in \mathbb{N} \land k \neq j \\
  2 \times \lfloor \log_2 n \rfloor & \text{Otherwise}
  \end{cases}
  \]

- $|E| = (n/2) \times \delta$

- Regular graph

- Vertex symmetric
Theoretical Properties of BMG

- Diameter \((D)\)
  \[
  O\left(\left\lceil \frac{\log_2(n)}{2} \right\rceil \right)
  \]

- Average Distance \((\bar{d})\)
  \[
  \approx \frac{\log_2(n)}{3}
  \]
Theoretical Properties of BMG

- Cost Factor ($\xi$) – Degree VS Diameter

\[ \xi = D \times \delta_{\text{max}} \]

\[ \xi = D \times |E| \]
Theoretical Properties of BMG

- Message Traffic Density ($\rho$)

\[
\rho = \frac{\bar{d} \times |V|}{|E|}
\]

$\rho_{BMG} \approx \frac{1}{3}$
Fault-Tolerant Properties of BMG

- Node-Connectivity ($\kappa$)
- Link-Connectivity ($\lambda$)
- Optimally Connected
  - $\kappa = \lambda = \delta_{\text{min}}$
Fault-Tolerant Properties of BMG

- Fault-Diameter ($F$)

- Strongly Resilient
  - $F(G) \leq D(G) + \Phi$

- Weakly Resilient
  - $F(G) \leq D(G) \times \Phi$
  - $F(BMG) \leq D(BMG) + 2$
Two-Terminal Routing Algorithms

- Breadth-First Search $O(\delta^D)$
- Basic-Greedy algorithm $O(\delta)$
  - Send to neighbor that has $ID$ closest to the $DestID$

- Estimate number of hops between $Src$ and $Dest$
  - Both Clockwise and Counter-Clockwise
  - Bit Counting $O(1)$ [e.g., distance 7 ($0111_2$) -> 4, 2, 1]
    - Parallel Bit Counting
    - Lookup Table
Two-Terminal Routing Algorithms

- Estimate number of hops between Src and Dest
  - Consecutive Bit Elimination $O(\log_2(n))$
    - CW Distance 7 ($0111_2$) -> CW 8($1000_2$), CCW 1($0001_2$)
    - Scan binary from right to left [ Skip bit-0 ]
    - Add value of least significant of cluster bits to both directions

  **Example** CW 110 ($1101110_2$) -> CW 110, CCW 0
  Add 2 ($10_2$) -> CW 112 ($1110000_2$), CCW 2($10_2$)
  Add 16 ($10000_2$) -> CW 128 ($10000000_2$), CCW 18 ($10010_2$)

- Equivalence Class $O(l \times \log_2(n))$
  - Distance $d$ is the same as $|d \pm (l \times n)|$
Two-Terminal Routing Algorithms

• Accuracy (for $n \leq 4096$)

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Overhead (%)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Maximum</td>
</tr>
<tr>
<td>BFS</td>
<td>00.00</td>
<td>00.00</td>
</tr>
<tr>
<td>Basic (Greedy)</td>
<td>05.55</td>
<td>11.38</td>
</tr>
<tr>
<td>Bit Count</td>
<td>25.13</td>
<td>36.23</td>
</tr>
<tr>
<td>Cons-Bit-Elimination</td>
<td>01.28</td>
<td>04.43</td>
</tr>
<tr>
<td>Equiv-Class (l=1)</td>
<td>00.62</td>
<td>02.47</td>
</tr>
<tr>
<td>Equiv-Class (l=2)</td>
<td>00.02</td>
<td>00.28</td>
</tr>
<tr>
<td>Equiv-Class (l=3)</td>
<td>7.07e-5</td>
<td>00.01</td>
</tr>
<tr>
<td>Equiv-Class (l=4)</td>
<td>00.00</td>
<td>00.00</td>
</tr>
</tbody>
</table>

• Performance

![Graph showing time (sec) vs. number of nodes for different algorithms]
Multicast (1 to $m$, $m < n$)

- Destinations are embedded in the header
  - Unlike IP multicast which requires join/leave group

- Based on two-terminal (unicast) routing
  - Messages can be split at an intermediate node
    - Split if the shortest path is in different directions
Broadcast (1 to \(n\)) / Gather (\(n\) to 1)

- Binomial spanning tree from source

- \(\lceil \log_2(n) \rceil\) steps
AllGather ($n$ to $n$)

- **Ring** [$n-1$ steps]: At step $s$
  - Node $i$ sends data of node $(i-s)$ to node $i+1$
  - Node $i$ receives data of node $(i-s-1)$ from node $i-1$

- **Gather + Broadcast** [$2 \log_2(n)$ steps]

- **Bruck** [$\log_2(n)$ steps]: At step $s$
  - Node $i$ sends data to node $i-2^s$
  - Node $i$ receives data to node $i+2^s$
Barrier (Synchronization)

- **Double - Ring** $[2n \text{ steps}]$:
  - Node exits from barrier when it receives a token second time

- **Gather + Broadcast** $[2 \log_2(n) \text{ steps}]$

- **Bruck** $[\log_2(n) \text{ steps}]$:
  - At step $s$
    - Node $i$ sends 0 byte message to node $i-2^s$
    - Node $i$ receives 0 byte message to node $i+2^s$
Routing in Failure Circumstances

- Unicast / Multicast / Ring-based
  - Reroute with alternate path
  - Prevent loop with transit list
- Broadcast / Gather
  - Broadcast message is encapsulated in a multicast sent from parent to children of the failed node.
- Bruck
  - Require additional multicast to get missing data
Reliability Analysis

- BMG
Reliability Analysis

- BMG vs. CST
Self-Healing BMG

- Adaptive
- Naïve VS. Adaptive
Self-Healing BMG

- Dynamic Expansion

![3D graph showing the number of connections between number of nodes and number of added nodes. The graph compares 'Adaptive' and 'Naive' methods.]
Multiple BMG

- Wide Area Communication
  - Multi-home is a node belonging to two or more groups
  - Gateway is a gateway of connection between groups

UUID = GroupID + NodeID

Messages will be automatically forwarded within adjacent groups
Conclusion

• Communication framework (library)
  • MPI implementation independent
  • New logical topologies
    • Graph theoretical (properties) analysis
      • Scalability and Fault-tolerance
    • Fault-tolerant routing protocols
      • Formally specified and verified protocols
    • Performance and reliability analysis
Future Work

• Aware physical topology
  • Function cost associates with links (Weighted graph)

• Reliability simulation
  • Support more distributions, e.g., LogNormal

• Generalized BMG
  • Node $i$ has links to \( \{i \pm 1, i \pm b, \ldots, i \pm b^k \mid b^k \leq n \} \)

• Integrate to MPI Implementation
  • BMG -> ORTE (Open Run-Time Environment)
  • RTST -> OPAL (Open Access Layer)